一种基于支持向量机的手写汉字识别方法

高 学[1],金连文[1],尹俊勋[1],黄建成[2]
(1. 华南理工大学电子与通信工程系,广东广州 510641;2. Motorola China Research Center,上海 200002)

摘 要: 本文提出了一种新的基于支持向量机手写汉字识别方法. 支持向量机作为一种新的机器学习方法,由于其建立在结构风险最小化准则之上,而不是仅仅使经验风险达到最小,从而使得支持向量分类器具有较好的推广能力. 本文首先讨论了支持向量机的基本原理,然后,针对支持向量机识别大类别手写汉字所遇到的特殊问题,文章进行了分析和阐述,并在此基础上,提出了基于最小距离分类器预分类的两级分类策略. 最后,针对 GB2312-80 的 1034 个汉字类别的 120 套手写样本,进行了实验仿真. 实验结果表明,本文方法的汉字识别率较距离分类器有较大提高,其中多项式核函数的支持向量分类器,识别率平均提高 3.38 %,表明了本文方法的有效性.

关键词: 支持向量机; 手写汉字识别; 特征提取
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2002) 05-0651-04

A New SVM-Based Handwritten Chinese Character Recognition Method

GAO Xue[1], JIN Lian-wen[1], YIN Jun-xun[1], HUANG Jian-cheng[2]
(1. *Dept. of Electronics and Communication Engineering, South China University of Technology, Guangzhou, Guangdong* 51064, *China*;
2. *Motorola China Research Center, Shanghai* 200002, *China*)

Abstract: A new recognition method of handwritten Chinese characters by support vector machine is presented. Support vector machines (SVM) operate on the principle of structure risk minimization which not only keeps the empirical risk minimal but also controls VC confidence of discriminant functions, hence a better generalization ability is guaranteed. In this paper, the problems to be solved while applying SVM in Chinese character recognition are addressed at first, and then a two stage of recognition scheme is suggested. Finally, experimental results on 1034 categories of Chinese character from 120 sets of samples are given. For the Polynomial kernel SVM, a 3.38 % average increment of recognition rate is obtained showing the efficiency of the proposed approach.

Key words: support vector machines; handwritten Chinese character recognition; feature extraction

1 引言

汉字识别一直是模式识别最重要的研究领域之一. 经过多年的研究,已经取得了大量成果[1~3]. 但是,无约束的非特定人手写汉字识别仍然被认为是文字识别领域最困难的问题之一,其原因可以归结为: (1) 汉字规模大 (2) 相似汉字较多,且有些相似字差别极其细微 (3) 存在大量的不规则书写变形. 由于 (2)、(3) 的存在,导致手写汉字,特别是相似字在特征空间中的距离变小,使得普通的距离分类器的推广能力变弱. 因此,如何补偿手写汉字的书写变形,提高分类器的泛化和推广能力,就成为汉字识别研究的关键问题之一. 针对汉字的结构特点,许多学者分别从预处理和特征提取的角度提出了许多方法[4~7]. 从预处理的角度,通过对汉字点阵采取某种非线性变换,矫正手写汉字变形,以减少类内方差. 从特征提取的角度,利用汉字固有的笔划构成特征提取手写汉字的笔划以及笔段信息. 但是,由于汉字的结构复杂性,以及不同人书写变形的不确定性,到目前为止,手写汉字的识别性能仍然不能令人满意. 神经网络由于其较强的曲线拟合和模式分类能力,在汉字识别中得到广泛的应用. 但是,神经网络方法也有其缺点,比如网络结构的确定尚无可靠的规则,算法的收敛速度较慢,且无法保证收敛到全局最有点. 本文提出了一种新的基于支持向量机的手写汉字两级分类策略,即采用最小距离分类器进行预分类,然后,应用支持向量机较强的泛化能力对候选字集进行细分类,取得了较好的效果. 支持向量机是 AT&T Bell 实验室的 V. Vapnik 等人根据统计学习理论提出的一种新的机器学习方法,已经在模式识别、回归分析和特征选择等方面得到了较好的效果[8,9]. 支持向量机可以看作一种新的训练多项式、径向基机分类器或神经网络分类器的方法. 根据结构风险最小化准则,在使训练样本分类误差极小化的前提下,尽量提高分类器的泛化推广能力. 从实施的角度,训练支持向量机等价于解一个线性约束的二次规划问题,使得分隔特征空间中两类模式点的两个超平面之间距离最大,而且它能保证得到的解为全局最优点,使得基于支持向量机的手写汉字

收稿日期:2001-06-29;修回日期:2001-12-03
基金项目:国家自然科学基金 (No. 69802007);广东省自然科学基金 (No. 980602);Motorola 研究基金

分类器能够吸收书写的变形,从而具有较好的泛化和推广能力.本文首先讨论了支持向量机的分类方法,然后给出了手写汉字的特征提取方法以及基于支持向量机的两级分类策略,最后针对 GB2312-80 的 1034 个汉字类别的 1034 个汉字的识别问题,进行了实验仿真.

2 支持向量机

首先考虑一个两类模式分类问题,设模式样本 (x_i, y_i) 服从空间 $X \times Y$ 上的某个未知概率分布 $P(x, y)$ (为方便起见,不妨令 $Y = \{1, -1\}$).目的是寻找一个分类函数,该分类函数将空间 $X \times Y$ 划分为两个子空间,不同的模式样本属于不同的子空间.在两类模式线性可分的情况下,超平面 (1) 可以将两类区分开.

$$w \cdot x + b = 0 \tag{1}$$

其中"·"表示向量的点积.

当两类模式为线性不可分的情况下,超平面 (1) 无法将两类分隔开,需要寻找某个非线性分类函数,这时可以通过一个非线性变换 $\Phi: x \to \phi(x)$,将给定模式样本变换到某个高维特征空间.然后,在高维特征空间中构造分类超平面.该分类超平面在原空间中可以表示为决策面

$$w \cdot \phi(x) + b = 0 \tag{2}$$

决策面应满足不等式约束

$$y_i(w_i \cdot \phi(x_i) + b) \geq 0, \quad i = 1, \cdots, n \tag{3}$$

其中 n 为训练样本数(以下同).

考虑到两类样本到超平面都应有一定的距离,约束条件 (3) 可以写为,

$$y_i(w_i \cdot \phi(x_i) + b) \geq 1, \quad i = 1, \cdots, n \tag{4}$$

显然,满足要求的分类超平面 (1) 和 (2) 不止一个,在结构风险最小化准则下,支持向量机寻找最优超平面,使得两类之间的间隔最大,同时保持训练样本的分类误差尽可能小.这里两类之间的间隔定义为两类模式样本到超平面的最近距离之和,如图 1 所示.

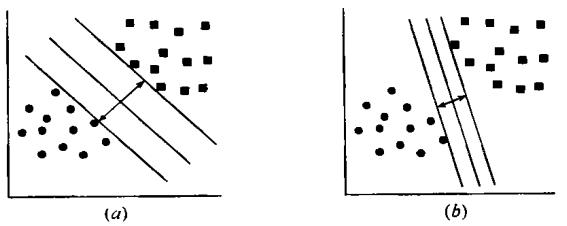

图 1 超平面位置示意图

(a) 间隔较大的分类超平面 (b) 间隔较小的分类超平面

寻找最优超平面可以归结为解如公式 (5) 所示的一个二次规划问题

$$\min\left\{\frac{1}{2} w \cdot w\right\} \tag{5}$$

约束条件

$$y_i(w_i \cdot \phi(x_i) + b) \geq 1, \quad i = 1, \cdots, n \tag{6}$$

然而,在许多实际应用中,完全满足式 (6) 的精确分类超平面是不存在的.因此,考虑到某些样本可能不满足约束条件 (6),引入松弛变量

$$\zeta_i \geq 0, \quad i = 1, \cdots, n \tag{7}$$

于是,约束条件 (6) 变为

$$y_i(w_i \cdot \phi(x_i) + b) \geq 1 - \zeta_i, \quad i = 1, \cdots, n \tag{8}$$

根据结构风险最小化准则,二次规划问题 (5) 变成 (9)

$$\min\left\{\frac{1}{2} w \cdot w + \sum_{i=1}^{n} \zeta_i\right\} \tag{9}$$

约束条件

$$y_i(w_i \cdot \phi(x_i) + b) \geq 1 - \zeta_i, \quad i = 1, \cdots, n \tag{10}$$

$$\zeta_i \geq 0, \quad i = 1, \cdots, n \tag{11}$$

其中 C 称为惩罚因子,通过改变惩罚因子可以在分类器的泛化能力和误分类率之间进行折衷.

为求解上述优化问题,引入拉格朗日函数 L

$$L(w, b, \zeta_i, \alpha_i, \lambda_i) = \frac{1}{2} w \cdot w + C \sum_{i=1}^{n} \zeta_i - \sum_{i=1}^{n} \lambda_i \zeta_i - \sum_{i=1}^{n} \alpha_i [y_i(w_i \cdot \phi(x_i) + b) - 1 + \zeta_i] \tag{12}$$

其中 $\alpha_i \geq 0, \lambda_i \geq 0$

函数 L 关于 w, b, ζ_i 最小化,同时关于 α_i, λ_i 最大化.根据极值存在的必要条件,函数 L 的极值应满足条件

$$\frac{\partial}{\partial w} L = 0, \frac{\partial}{\partial b} L = 0, \frac{\partial}{\partial \zeta_i} L = 0 \tag{13}$$

解方程 (13) 并代入 (12),可以得到优化问题 (9) 的对偶形式,即最大化函数

$$w(\alpha_i) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{14}$$

约束条件

$$0 \leq \alpha_i \leq C, \quad i = 1, \cdots n \tag{15}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{16}$$

以及超平面的系数向量

$$w = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i) \tag{17}$$

其中 $K(x_i, x_j = \Phi(x_i) \cdot \phi(x_j)$ 称作核函数.

解优化问题 (14) 可以得到系数 α_i,代入 (17) 则可以确定超平面系数向量 w.于是分类函数可以表示为

$$f(x) = \operatorname{sgn}\left[\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b\right] \tag{18}$$

可以证明,只有一部分 (通常是少部分) α_i 不为零,其相应的样本称作支持向量.且对于 $\alpha_i \in (0, C)$,其支持向量满足 (19)

$$y_i = (w_i \cdot \phi(x_i) + b) = 1 \tag{19}$$

从而可以确定分类函数 (18) 中的常数 b.

从 (14)、(18) 可以看出,尽管支持向量机通过一个非线性变换,把线性不可分的问题转化为高维空间中的线性问题,无论是优化函数 (14) 或者分类函数 (18) 都只涉及到样本之间在高维空间的内积运算,而这种内积运算可以用原空间中的函数实现.根据泛函的有关理论,只要核函数 $K(x_i, x_j)$ 满足 Mercer 条件,它就对应某一变换空间中的内积.因此,通过选择合适的内积函数,就可以实现某一非线性变换后的线性可分.

对于 m 类模式的分类问题,可以设计 m 个两类分类器,

每个分类器只区分一类模式与其它类. 给定输入模式 x,设 m 个分类函数为

$$f^j(x) = \sum_{i=1}^{n} \alpha_i^j y_i K(x, x_i) + b^j, \quad j = 1, \cdots, m \tag{20}$$

在理想情况下,应存在某个 $k \in \{1, \cdots, m\}$,使得

$$f^k(x) = \max_{j=1,\cdots,m} f^j(x) > 0 \tag{21}$$

且满足

$$f^j(x) < 0, \quad j = 1, \cdots, k-1, k+1, \cdots, m \tag{22}$$

则输入模式应属于第 k 类. 为了增加分类器输出的可靠性,可以采用更严格的判据条件. 即,如果

$$f^k(x) > \tau \tag{23}$$

$$f^j(x) < -\tau, \quad j = 1, \cdots, k-1, k+1, \cdots, m \tag{24}$$

其中 $\tau > 0$.

则判定输入模式应属于第 k 类.

3　特征提取

特征提取是手写体汉字识别研究的另一个重要问题,所提取特征的好坏将直接影响整个汉字系统的性能. 文献[10]尝试了应用支持向量机进行英文字母的识别,文中直接采用字母图象的象素值作为特征,尽管也取得了较好的效果,但是,对于大类别的汉字识别系统将是不可取的. 首先,使用象素值作为特征将使特征的维数增加,增加运算和存储量. 例如,64 ×64 的汉字图象,其特征维数将达到 4096. 其次,由于汉字本身固有的结构特点,即汉字主要有横、竖、撇、捺四种笔划组成. 有效的提取汉字的笔划信息不仅会降低特征维数,而且会提高系统的识别性能. 本文提取基于点密度的弹性网格方向特征[11]作为支持向量机的输入样本. 汉字图象经二值化和归一化后,根据汉字图象的黑象素点的密度分布划分一定大小的弹性网格. 然后,汉字图象经过细化,根据 8 邻域象素点分布情况,把每个黑象素点分解到横、竖、撇、捺四种方向子模式. 最后,根据四种方向子模式的每个弹性网格中的笔划象素点的分布情况计算网格方向特征. 对于 8 ×8 的网格,则特征维数降为 256. 图 2 为"哎"、"安"的网格划分和方向分解结果.

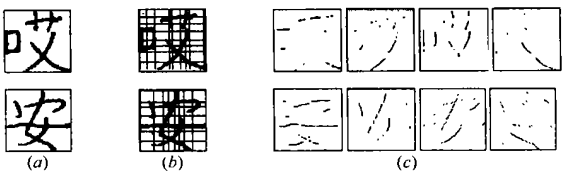

图 2　(a)归一化后的汉字点阵　(b)8 ×8 弹性网格　(c)四种方向分解子模式.

4　基于支持向量机的两级分类策略

应用支持向量机进行大类别的手写汉字识别,无论是训练或识别阶段,减少存储和计算量都是一个重要问题. 例如,选择 50 套手写汉字样本,每套取 1034 个汉字作为训练样本,每个汉字提取 256 维的特征,如果每个字都选用所有的样本进行支持向量机的训练,则每个支持向量机都有 51700 ×256 个训练数据. 那么支持向量机的训练将花费很多时间(在本文的实验中,训练支持向量机算法采用 JohnC. Platt[12]提出的序列最小化优化方法). 同时,训练样本的增加也会导致支持向量数目的增加. 基于以上分析,本文提出基于距离分类器预分类的两级分类策略,支持向量机的训练仅采用距离分类器给出的候选字集作为训练样本,这可以使计算和存储量大大降低,同时支持向量的数目也相应减少. 而且,利用距离分类器的输出信息可以改善支持向量分类器的性能. 本文的距离分类器采用最小欧氏距离分类器,其前 n 个候选字的识别性能如表 1 所示.

表 1　最小欧氏距离分类器的前 n 个候选字的识别率(1034 个汉字规模)

候选字数	5	10	15	25	50
识别率(%)	97.87	98.62	98.94	99.31	99.58

在支持向量机训练阶段,设给定输入模式 x,距离分类器给出的 n 个候选字为 $C_1, \cdots, C_n$. 如果 $x \in C_i, i = 1, \cdots n$,则 n 作为训练 C_i 的支持向量机的正样本,否则为负训练样本.

在汉字的识别阶段,设 $C_1, \cdots, C_n$ 为距离分类器给出的未知样本 x 的 n 个候选字类别,其相应的距离度量值为 $d_1, \cdots, d_n$,且由小到大排列,$f^1(x), \cdots, f^n(x)$ 为相应支持向量分类器的输出.

$$d_i = \| x - x_i^T \|^2, \quad i = 1, 2, \cdots, n \tag{25}$$

其中 x_i^T 为距离分类器第 i 类汉字的模板向量.

令

$$f_0 = f^k(x) = \max_{i \in \{1, \cdots n\}} f^i(x) \tag{26}$$

$$f_1 = \max_{i \in \{1, \cdots n\}, j \neq k} f^j(x) \tag{27}$$

其中 $k \in \{1, 2, \cdots n\}$.

支持向量机分类器采用以下的加权分类策略:

(1)如果 $f_0 \geq \tau$ 且 $f_1 \leq -\tau$(τ 为正的常数),则未知样本属于 C^k 类. 否则转(2)

(2)分别将距离分类器以及支持向量分类器的输出 $d_1, \cdots, d_n$ 和 $f^1(x), \cdots, f^n(x)$ 进行归一化到 $[0, 1]$,并计算加权输出. 设 $\tilde{d}_1, \cdots, \tilde{d}_n$ 和 $\tilde{f}^1(x), \cdots, \tilde{f}^n(x)$ 为归一化结果,综合考虑距离分类器和支持向量分类器的分类能力,则取加权输出为

$$f_w^i(x) = -w_1 \tilde{d}_i + w_2 \tilde{f}^i(x), \quad i = 1, \cdots, n \tag{28}$$

其中 w_1 和 w_2 分别为距离分类器以及支持向量分类器的加权系数,分别与相应分类器的识别性能成正比.

如果

$$f_w^k(x) = \max_{i \in \{1, \cdots, n\}} f_w^i(x) \tag{29}$$

则未知样本属于 C^k 类.

5　实验结果

为了检验本文提出的手写汉字识别方法的有效性,随机地从 863 手写体汉字字库 HCL2000 中取出 120 套样本,每套样本取国标 16 - 26 区的 1034 个汉字. 其中 100 套样本用于确定距离分类器的模板以及支持向量分类器的训练,其余 20 套用于系统的识别性能测试,每个汉字样本划分为 8 ×8 的弹性

网格并提取 256 维的特征向量. 目前研究最多的支持向量机核函数主要有三类:

(1) 多项式核函数

$$K(x, x_i) = [(x, x_i) + 1]^d$$

(2) 径向基核函数

$$K(x, x_i) = \exp\left\{\frac{|x - x_i|^2}{\sigma^2}\right\}$$

(3) Sigmoid 核函数

$$K(x, x_i) = \tanh(v(x \cdot x_i) - \theta)$$

其中 d, σ, v, θ 均为常数.

本文分别就三种不同的核函数对所提出方法进行了识别性能测试,结果如表 2 所示.

表 2 三种不同核函数的支持向量机 1034 个手写汉字的识别率(%)

τ	距离分类器	0	0.1	0.2	0.3	0.4
多项式核函数 ($d=3$)	91.92	95.42	95.35	95.26	95.25	95.24
径向基核函数	91.92	94.47	94.55	94.56	94.59	94.57
Sigmoid 核函数 ($\theta=0.5$)	91.92	94.44	94.45	94.43	94.42	94.41

(其中 τ 为加权分类策略中的常数.)

从表 2 可以看出,三种核函数的支持向量分类器的识别率均较距离分类器有较大幅度的提高,其中尤以多项式核函数的分类器识别率最高,平均提高 3.38%,表明本文提出的方法是行之有效的. 同时也可以看出,多项式核函数的识别率较其它两种核函数分类器的识别率略有提高,尽管提高的幅度还不足以表明,多项式核函数的支持向量机分类器一定优于其它两种分类器,但至少可以得出结论,对于具体的应用问题,不同的核函数的分类性能是有差异的.

6 讨论

非特定人手写汉字识别被认为是文字识别领域最为困难的问题之一. 为了解决这个问题,人们分别从汉字的预处理、特征提取与选择、分类器设计等不同的角度提出了许多方法,但到目前为止,非特定人的手写体汉字识别仍然缺乏行之有效的方法. 本文从分类器的角度,提出了一种基于支持向量机的手写汉字识别方法. 支持向量机作为一种新的机器学习方法,由于其建立在结构风险最小化准则上,而不仅仅是经验风险最小,从而使得其具有较强的泛化推广能力. 本文尝试利用支持向量机较强的泛化能力,解决相似汉字以及笔划书写的不规则变形带来的识别困难. 从实验结果看,汉字识别率有较大的提高,表明支持向量机方法在手写汉字的识别领域具有较好的应用前景. 同时,支持向量机核函数的选择对分类器的识别性能也有影响,如何利用关于汉字的先验知识,选择和设计合适的核函数,还有待进一步研究.

参考文献:

[1] Hildebrand T H, Liu W. Optical recognition of handwritten Chinese characters: advances since 1980 [J]. Pattern Recognition, 1993, 26 (2): 205 - 225.

[2] Tsukumo J. Handprinted Kanji OCR development - what was solved in handprinted Kanji character recognition [J]. IEICE Trans Inf And Syst. 1996, E79-D: 411 - 416.

[3] Kimura Y, Wakahara T. Toward robust handwritten Kanji character recognition [J]. Pattern Recognition Letters, 1999, 20: 979 - 990.

[4] 陈友斌, 丁晓青, 吴佑寿. 一种新的用于手写汉字识别的非线性归一化方法 [J]. 模式识别与人工智能, 1998, 11(3): 310 - 317.

[5] Lee C N, Wu B H. A chinese character stroke extraction algorithm based on contour information [J]. Pattern Recognition, 1998, 31(6): 651 - 663.

[6] Lin J R, Chen C F. Stroke extraction for Chinese characters using a trend following transcribing technique [J]. Pattern Recognition, 1996, 29(11): 1789 - 1803.

[7] 孙星明, 等. 完全基于结构知识到汉字笔划提取方法 [J]. 计算机研究与发展, 2000, 37(5): 543 - 550.

[8] Vapnik V. The Nature of Statistical Learning Theory [M]. 张学工, 译. Spring-Verlag New York Inc, 1995. 统计学习理论的本质 [M]. 北京: 清华大学出版社, 1999.

[9] Burges J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 1 - 47.

[10] Schomaker L, Vuurpijl L. Support vector machines for the classification of western handwritten capitals [A]. Proceedings of the Seventh International Workshop on Frontiers in Handwritting Recognition [C]. 2000. 167 - 176.

[11] Jin Lianwen, Wei Gang. Handwritten Chinese character recognition with directional decomposition cellular features [J]. Journal of Circuits, System, and Computers, 1998, 8(4): 517 - 524.

[12] Platt J C. Sequential minimal optimization: a fast algorithm for training support vector machines. advances in kernel methods-support vector learning [A]. Cambridge, MA: MIT Press [C]. 1999. 185 - 208.

作者简介:

高 学 男, 1967 年 9 月出生于河南省西华县, 华南理工大学电子与通信工程系 2000 级博士生, 主要研究领域: 汉字识别, 图象处理, 遗传算法.

金连文 男, 1968 年 10 月出生于贵州省都匀, 华南理工大学电子与通信工程系副教授, 博士, IEEE 会员, 主要研究领域: 中文信息处理, 神经网格, 模式识别, ASIC 设计.

尹俊勋 男, 1942 年生于广东省东莞, 现为华南理工大学电子与通信工程系教授, 博士生导师, 主要研究领域: 音频与视频信号处理, 模式识别, CDMA.